Proceedings of the 28th IEEE
EMBS Annual International Conference
New York City, USA, Aug 30-Sept 3, 2006

Syndrome-Based Discrimination of Single Nucleotide Polymorphism

E. E. May[1], P. Dolan[2], P. Crozier[3], S. Brozik[2]
[1]Computational Biology Department
[2]Biosensors and Nanomaterials Department
[3]Multiscale Computational Material and Methods Department
Sandia National Laboratories, Albuquerque, NM 87185 USA
(Email: eemay@sandia.gov)

Abstract— **The ability to discriminate nucleic acid sequences is necessary for a wide variety of applications: high throughput screening, distinguishing genetically modified organisms (GMOs), molecular computing, differentiating biological markers, fingerprinting a specific sensor response for complex systems, etc. Hybridization-based target recognition and discrimination is central to the operation of nucleic acid sensor systems. Therefore developing a quantitative correlation between mishybridization events and sensor output is critical to the accurate interpretation of results.**

In this work, using experimental data produced by introducing single mutations (single nucleotide polymorphisms, SNPs) in the probe sequence of computational catalytic molecular beacons (deoxyribozyme gates) [1], we investigate coding theory algorithms for uniquely categorizing SNPs based on the calculation of syndromes.

Keywords— **DNA sensors, hybridization, SNP, single nucleotide polymorphism, coding theory**

I. Introduction

Hybridization is a fundamental event that controls numerous *in vivo* biological processes from the regulation and production of proteins to the response of the immune system to foreign and self-antigens [2]. In the genetic process of protein translation, the correlation between mRNA/tRNA interactions and amino acid recruitment is well captured in the genetic code, which maps RNA codons to their amino-acid counterparts. Uncovering this fundamental concept has had immeasurable consequences for molecular biology. Understanding the causative relation between hybridization and biological events such as transcription and translation regulation or *in vitro* events such as detectable fluorescence signatures is a quantitative challenge.

Polymorphisms are the mediators of diverse phenotypic expressions in biological organisms. A single nucleotide polymorphism (SNP) in one nucleotide position may differ phenotypically than a SNP in another. This phenotypic disparity can be observed in natural systems and in nucleotide-based sensor systems. We have observed that a single nucleotide polymorphism (SNP) in one position of our input DNA probe differs phenotypically from a SNP in another, causing the activated deoxyribozyme to produce distinct fluorescence signatures, Figure 1 [1]. In Figure 1 the horizontal axis represents time and the vertical axis represents fluorescence.

Fluorescence signatures were measured with FluoDia T70 Microplate Reader (PTI, Inc.) using black 384-well microplates (OptiPlate-384F, Perkin-Elmer) at room temperature. Each well measured represents a 55 μl detection volume containing: i) Reaction Buffer, ii) the deoxyribozyme molecular beacon, iii) fluoregenic substrate, and iv) a 15-base DNA input probe sequence. Figure 1 shows the fluorescence output for the non-SNP input sequence and the 45 SNPs associated with the input sequence.

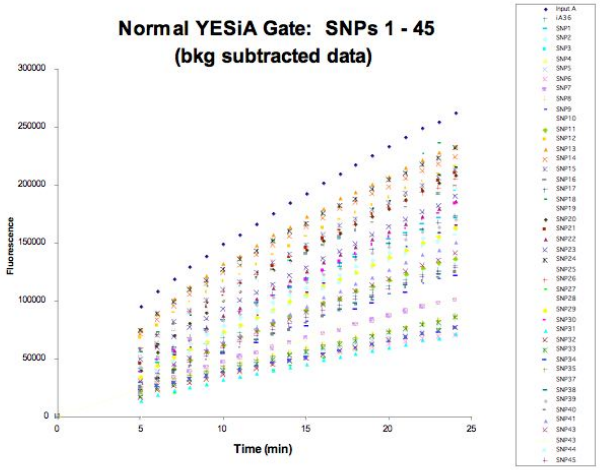

Fig. 1. SNP detection using indicator ribozyme gate: Differential hybridization produces different fluorescence signatures.

When there is a single base mishybridization between the input probe sequence and the recognition loop of the deoxyribozyme, there is a change in the the thermodynamic stability of the system. Quantitatively this change can be assessed using measures such as free energy and melting temperature. In related work, we have found significant correlation between empirically calculated free energy and melting temperature values of the target/probe pair and observed fluorescence values. While it is relatively easy to distinguish a perfect hybridization event from a mishybridization event, further classification of mishybridization events into location and type of SNP presents a greater challenge.

The accuracy of a hybridization-based biosensors depends on how well we are able to determine if the target DNA or a mutated form of the target is encountered based on the fluorescent signature. Another way of thinking of this is how to classify sequences into a correct or error-containing group. Using ideas from the field of coding theory, we can draw parallels between detection and classification of mishybridizations and detection and classification of error vectors for error control codes [3].

II. Methods

A. ECC and Syndrome Generation

Error control codes (ECC) are algorithms that recognize a set of target signals (binary bit streams in digital communications) and specific variations (errors) of the reference signal set [4]. The reference signal set is called a codebook, C, and is a matrix where each row corresponds to a codeword vector. The ECC is usually represented by a generator matrix, G, which has a corresponding dual matrix H, referred to as the parity check matrix. The relationship between the parity check matrix and the codebook is: $S = H^T * C$, where (*) represents multiplication, H^T is the transpose of the parity check matrix, S is the set of syndrome values, and S equals zero if there are no errors in the sequences in C.

Each vector, s, in S represents an error vector. Unique errors in our codewords should produce unique error vectors until the number of errors exceeds the error correction limit of our error control code. The transpose of H is our syndrome generation algorithm; this can be used to detect and classify variations in C . If the codebook, C , represents the sequence group we want to recognize, then our overall goal is to develop methods for finding H^T.

B. Syndrome Generator for SNPs

Using the sequence data from Figure 1, we reverse-engineered an ECC syndrome generation algorithm that can identify a non-SNP target and SNPs of the target using computed syndromes. We pose the problem as follows: Find H^T_{YESiA} such that

$$H^T_{YESiA} * C_{YESiA} = S_{YESiA} = Zero \tag{1}$$

where

- C_{YESiA} is composed of the correct sequence and all SNP variations of this sequence.
- H^T_{YESiA} is in systematic form [4], [5]. For systematic (n, k) codes, G and H are of the form

$$G \ = \ [I_k; P] \tag{2}$$

$$H \ = \ [P^T; I_{n-k}] \tag{3}$$

where P is a k *by* $(n-k)$ matrix and I represents the k *by* k (or $(n-k)$ *by* $(n-k)$) identity matrix [4], [6]. Assuming a systematic code reduces the number of unknowns in the H matrix by $(n-k)^2$. The systematic form also simplifies conversion from H back to G.
- The best H^T_{YESiA} maximizes the number of zeros in S_{YESiA} while minimizing the number of zeros in H^T_{YESiA} [4], [5].
- We assume a ($n = 15$, $k = 5$) code. This coding rate is equivalent to a rate $\frac{1}{3}$ code, similar to the degeneracy of the codon to amino acid genetic code.

We convert the input DNA sequence set (InputA plus 45 SNPs) into their binary equivalence using chemical activity-based mapping proposed by MacDonaill (MacDonaill 2002)[7]. The binary coding rate is $(n = 60, k = 20)$, where each DNA base is represented by a four bit binary sequence. The binary sequence set serves as inputs to a genetic algorithm based error control code inverter for linear block codes [5]. Using the resulting H^T_{YESiA} we calculate the syndrome for each of the 46 input sequences in Matlab.

III. Results and Discussion

The resulting synrome generation algorithm, H^T_{YESiA}, had a fitness of 0.8609, where 1.0 is the maximum fitness. The optimal solution produces an H^T that optimizes a cost function of the form:

$$Fitness \ = \ R_S \frac{|Zeros \ \ in \ \ S|}{|S|} + R_P \frac{|Nonzeros \ \ in \ \ P|}{|P|} \tag{4}$$

where S represents the syndrome matrix (each row in S corresponds to the syndrome of a codeword in the codebook C) and $R_S + R_P = 1.0$. For this work R_S and R_P are 0.70 and 0.30, respectively. Our fitness score of 0.8609 indicates the resulting H^T_{YESiA} maximizes the number of zero elements in the syndrome matrix while minimizing the number of non-zeros in the parity submatrix of H^T_{YESiA}. Using this code we calculate the syndrome for each of the 46 input sequences. Figure 2 to Figure 4 show examples of the resulting syndrome vectors. The horizontal axis is position in the syndrome vector and the vertical axis is the syndrome value at each position.

The non-SNP input (InputA) produced an all zero syndrome, which in the coding theoretic sense indicates the absence of errors. The sequences with SNPs near the middle through 3' ends produced unique syndrome vectors (Figures 3 and 4, corresponding to SNPs in position six to fifteen). Because of their unique syndrome patterns these syndrome vectors can be used to uniquely identify the SNPs in the middle to 3' end and possibly the type of mutation that occurred. Usually the transition mutations have more ones in their syndrome than transversion mutations. Syndrome vectors corresponding to SNPs near the 5' end (Figure 2) contain a greater number of non-zero syndrome bits and their resulting syndrome patterns are not easily distinguished. Thus it may not be as easy to use these syndrome vectors to uniquely locate and classify the type of SNP that occurred in the input sequence.

We analyzed the syndrome vectors based on their Hamming distance. Figure 5 shows a contour of these distance values. SNPs that produced unique syndrome vectors, in general, clustered around SNP base locations. If we omit the non-SNP sequence, the minimum Hamming distances correspond to SNPs that are co-located. Conversely, SNPs in position one to five do not cluster according to SNP location.

IV. Conclusion

The syndrome generation algorithm can be used to uniquely distinguish SNPs in the middle to 3' end of the 15-base input sequence. This error control coding algorithm not only detects a mutation in the sequence but can categorize the location and type of base mutation based on the unique syndrome vectors. The question remains if there is a complementary algorithm that produces similar results for SNPs in the 5' end of the input sequence.

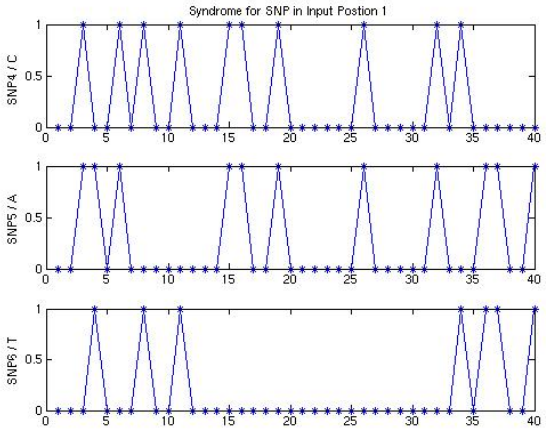

Fig. 2. *Syndrome for SNP in base position 1.*

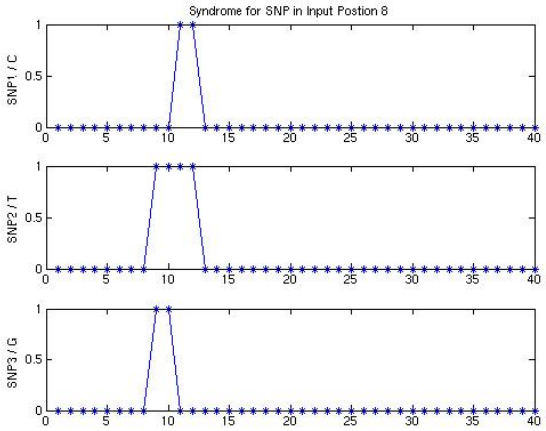

Fig. 3. *Syndrome for SNP in base position 8.*

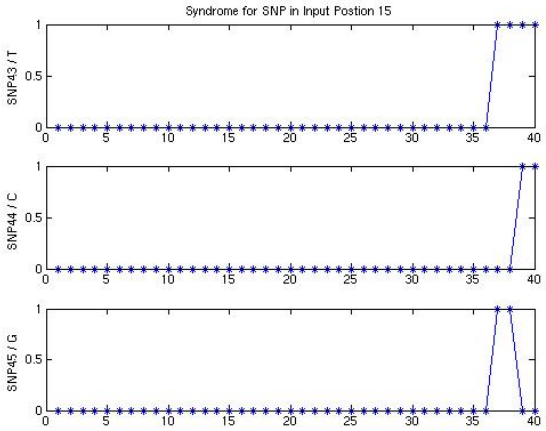

Fig. 4. *Syndrome for SNP in base position 15.* Unique syndrome pattern can be used to identify location and type of genetic modification. (Figure 2 to 4) Syndromes from SNPs on the 5, middle, and 3 ends of the input probe sequence.

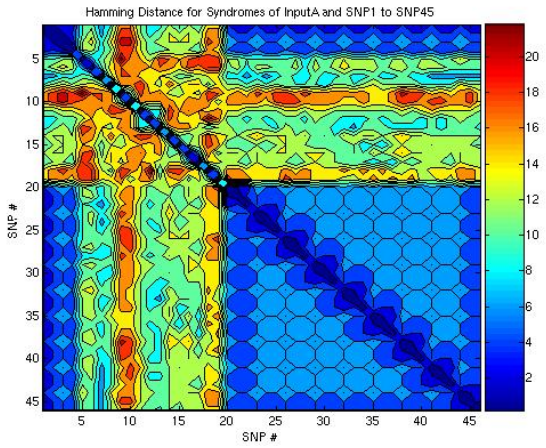

Fig. 5. Contour map of the Hamming distance of the non-SNP and SNP syndrome values. (Blue corresponds to lower distance values and red to higher distance values.)

Another remaining question is the issue of biological significance of the resulting code. We are using the approach discussed to evaluate whether a small subset of SNPs can be used to uniquely identify the occurrence and location of a larger set of SNPs.

In addition to coding theoretic methods, we also investigated and compared whether a statistical method can identify type or/and location of SNPs using the fluorescence signatures from Figure 1. We constructed a Bayesian classifier based on the output fluorescence data at a single time point. For the two class system, we classify an output fluorescence as either belonging to to a SNP in the upstream region (position 1 to 7) or the downstream region (positions 8 to 15). The correct classification rate for this classifier was 71.1%. The second system classifies a fluorescence as belonging to the 5' end (base position 1 to base position 5), the middle (base position 6 to base position 10), or the 3' end (base position 11 to 1base position 5). The correct classification rate for this classification system was 62.2%.

Although we are confident that better statistical classifiers can be constructed, the syndrome generation algorithm provides a non-statistical approach for correlating hybridization events with phenotypic events that occur both *in vitro* and *in vivo*.

Acknowledgments

Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company for the United States Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

References

[1] M. N. Stojanovic and D. Stefanovic, "A deoxyribozyme-based molecular automaton.," *Nature Biotechnology*, vol. 21, pp. 1069–107, 2003.

[2] Benjamin Lewin, *Genes V*, Oxford University Press, New York, NY, 1995.

[3] Elebeoba E. May, Mladen A. Vouk, Donald L. Bitzer, and David I. Rosnick, "Coding theory based models for protein translation initiation in prokaryotic organisms," *BioSystems*, 2004.

[4] Shu Lin and Daniel J. Costello, *Error Control Coding: 2nd Edition*, Prentice-Hall, Inc., Englewood Cliffs, NJ, 2004.
[5] Elebeoba E. May, "Optimal Generators for a Systematic Block Code Model of Prokaryotic Translation Initiation. ," in *25th Silver Anniversary International Conference of the IEEE Engineering in Medicine and Biology Society*, 2003.
[6] John B. Anderson and Seshadri Mohan, *Source and Channel Coding An Algorithmic Approach*, Kluwer Academic Publishers, Boston, MA, 1991.
[7] D. MacDonaill, "A Parity Code Interpretation of Nucleotide Alphabet Composition," *Chem Communic*, pp. 2062–2063, 2002.